



Analysis and Modelling of False Positives in GPCR Assays

Dipan Ghosh¹✉, Igor Tetko² , Bert Klebl¹, Peter Nussbaumer¹,
and Uwe Koch¹

¹ Lead Discovery Center GmbH, Otto-Hahn-Straße 15,
44227 Dortmund, Germany

{ghosh, klebl, nussbaumer, koch}@lead-discovery.de

² Institute of Structural Biology, Helmholtz Zentrum München – German
Research Center for Environmental Health (GmbH), Ingolstaedter Landstrasse 1,
85764 Neuherberg, Germany
i.tetko@helmholtz-muenchen.de

Abstract. G-Protein Coupled Receptors (GPCR) are involved in all the major signaling pathways. As a result, they often serve as potential target for therapeutic drugs. In this study we analyze publicly available assays involving different classes of GPCR to identify false positives. Using the latest developments in Machine Learning, we then build models that can predict such compounds with high confidence. Given the ubiquity of GPCR assays, we believe such models will be very helpful in flagging potential false positives for further testing.

Keywords: Neural Networks · Least squares SVM · Frequent hitters

1 Introduction

G-Protein Coupled Receptors (GPCR) are the largest family of cell surface receptors [1]. These plasma membrane bound receptors have evolved to recognize a variety of extracellular physical and chemical signals and, upon recognition, act as the proximal stimulus in cell signaling pathways. With over ~ 800 members [2], GPCRs are involved in almost every physiological function, from sensation to growth to hormone responses. Due to their widespread physiological relevance and presence of druggable sites, GPCRs are one of the major targets of therapeutic drugs. A 2017 study notes that 475 drugs act at 108 unique GPCRs. Approximately 321 agents are currently in clinical trials, of which $\sim 20\%$ target 66 potentially novel GPCR targets. GPCRs also account for $\sim 27\%$ of the global market share of therapeutic drugs, with aggregated sales for 2011–2015 of \sim US\$890 billion.

As promising drug targets, assays involving a member of the GPCR family are commonly employed in high throughput screening (HTS) campaigns. There are a plethora of different techniques and a wide range of commercial kits available, many of which are suitable for High Throughput Screening (HTS) [3]. In such HTS, identifying false positives is a challenge. False positives may be compounds that interfere with the assay detection technology in some way, such as inhibiting luciferase in luciferase-

based system [4], or quenching fluorescence where it is the final readout [5]. There may also be compounds that are not specific to the target protein, but are promiscuous, either to a narrow or broad class of proteins [6].

In the previous study we developed a machine learning method to flag potential frequent hitters for luciferase assays [4]. In this study we investigated whether the developed methodology can be extended to identify false positives for GPCR assays.

2 Data

2.1 Data Description

Our initial goal was to explore the available data and find suitable assays that we can then use for further analysis. On PUBCHEM, we identified 92 assays with more than 500 compounds for GPCR agonists and antagonists. We separated the two and decided to focus on the agonists. This was just to narrow down the scope of the study. From the list of available agonist screenings we selected the 20 assays with the highest number of active compounds. This is because as we are looking for false positives. Assays that have little to no positives are less relevant for us. For further selection particular assays, we focused on the GPCR subtypes as described below.

2.2 Data Collection

The GPCR family is commonly classified into five different families based on their structural and sequence similarity. The families are then further classified into a family tree [7, 8]. Of these five major families, the Rhodopsin class is the largest. For selecting assays for our analysis, we mapped the target proteins onto this family tree (Fig. 1), and selected assays with set of representative proteins distant from each other in the family tree. This ensures that compounds that are frequently active, are not preferential agonists of a subtype of GPCR, but are more likely a result of assay artifact.

Using these criteria we chose a set of 12 assays and looked for compounds that are frequently active in these assays (see Methods section), i.e. actives across all of the various different subtypes and assay technologies and thus frequent hitters of the Rhodopsin class of GPCR. However, only 59 out of 373,131 compounds matched our definition of being frequently active. Upon closer examination we found that these compounds were tested only thrice, and therefore are more likely to be an artifact of selection criteria rather than a GPCR frequent hitter or assay artifact.

To further refine our search, we next focused on different detection technologies that were used in the assays. We found that half of assays (six) used fluorescence while other six assays used bioluminescence. Only 71 compounds were frequently active in the bioluminescence group. In the fluorescence group, although the number of data-points and active compounds was very similar, 502 compounds were frequently active (Table 1). This indicates that fluorescence technology contributes many more artifacts and these 502 compounds were used for further analysis.

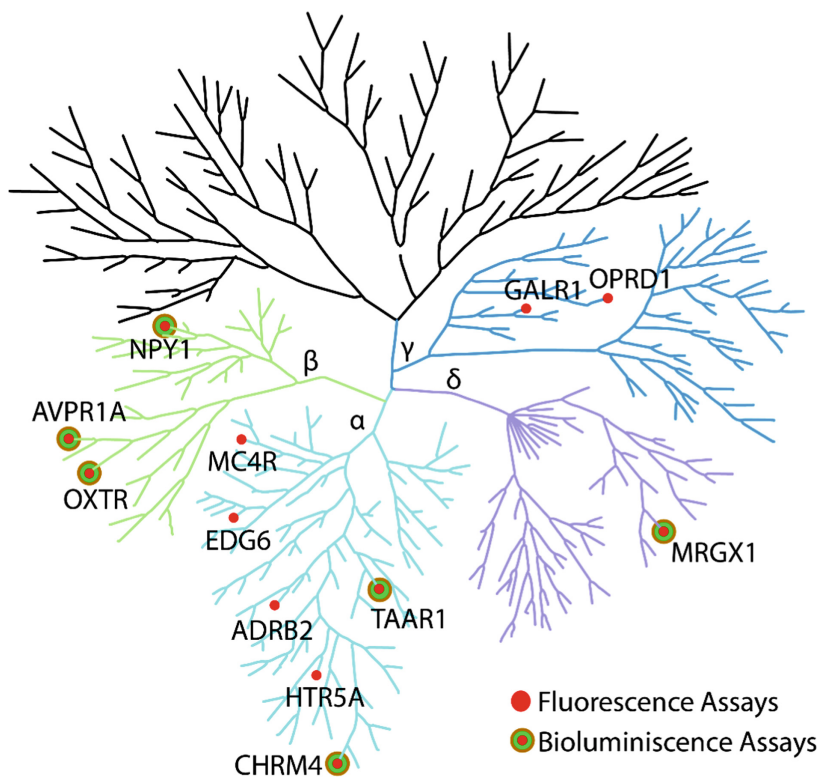


Fig. 1. GPCR family tree represented as a tree and dots mapping the protein targets in identified assays. The colored part of the tree represents the Rhodopsin class of the GPCR family and various subfamilies of the Rhodopsin class are marked with different colors. (Color figure online)

All data were harvested from PUBCHEM [9], manually or by using the PUBCHEM REST API with Python. All data were obtained and stored locally in the CSV format to be analyzed later with various python scripts.

Table 1. Statistics of compounds for the datasets used in the study.

	Inactive	Active	Frequently active
All assays	352685	20446	59
Fluorescence assays	363459	9605	502
Bioluminescence assays	358770	10841	71

2.3 Frequent Hitter Flagging

We defined frequent hitters as compounds that were active according to our criteria in more than half of the assays they were tested in. Additionally each compound had to be tested at least in three different assays. Compounds satisfying both criteria were identified using a Python script and flagged as frequent hitters.

3 Methods

Using the freely accessible platform On-line Chemical and Modeling Environment (OCHEM) [10], was used to build models for our data. Different descriptors available in OCHEM include CDK, Dragon 6 and 7, ISHIDA fragmentor, among others. Their detailed description can be found elsewhere [4]. Associative Neural Networks (ASNN) [11], Deep Neural Network (DNN) [12], Extreme Gradient Boost (XGBOOST) [13], and Least Squares Support Vector Machine (LSSVM) [14] algorithms were analyzed for training the models. The methods were used with default parameters as specified on the OCHEM web site.

4 Results and Discussion

4.1 Machine Learning

The analyzed methods were used in combination with different descriptors sets. LSSVM provided on average the highest accuracy amid the chosen algorithms (Table 2). We selected LSSVM models with the highest accuracy based on their ROC-AUC score for building a consensus model. The consensus model had ROC-AUC score of 0.93 with balanced accuracy of 86%.

To test our model, we constructed an independent dataset by looking up GPCR agonist assays in PUBCHEM that we did not use for the training set. We found five relevant assays with 4323 active compounds. Our frequent hitter analysis identified 157 compounds from these 5 assays. Our consensus model predicted the molecules from this set with a balanced accuracy of 76% and an AUC score of 0.85. The consensus model which was based only subset of 2D descriptors provided a very similar accuracy of 75% and AUC score of 0.85 thus indicating the importance of only 2D information for this analysis.

Table 2. The performance of models built using the GPCR dataset. The ROC-AUC scores are calculated using 5-fold stratified cross-validation. Models marked with asterisk were used to build the consensus model.

Descriptors/methods	DNN	ASNN	XGBOOST	LSSVM
ALogPS, OEGate (2D)	0.84	0.84	0.87	0.89*
CDK2 (3D)	0.79	0.85	0.86	0.87*
ChemaxonDescriptors (3D)	0.82	0.82	0.84	0.88*
Dragon6 (2D blocks)	0.83	0.87	0.88	0.91
Dragon6 (3D, all blocks)	0.87	0.85	0.89	0.91*
Fragmentor (2D)	0.85	0.83	0.88	0.89*
GSFrag (2D)	0.81	0.8	0.86	0.85
InductiveDescriptors (3D)	0.79	0.78	0.79	0.83
JPlogP (2D)	0.82	0.79	0.85	0.84
Mera, Mersy (3D)	0.69	0.76	0.8	0.81
PyDescriptor (3D)	0.89	0.86	0.85	0.89*
QNPR (2D)	0.81	0.82	0.87	0.86
RDKit (2D, all blocks)	0.88	0.88	0.87	0.91*
RDKit (3D, all blocks)	0.88	0.88	0.87	0.91
SIRMS (2D)	0.86	0.83	0.86	0.87*
Spectrophores (3D)	0.63	0.69	0.72	0.68
StructuralAlerts (2D)	0.79	0.79	0.77	0.78
alvaDesc (2D blocks)	0.86	0.85	0.87	0.91*
alvaDesc (3D, all blocks)	0.88	0.86	0.88	0.91

5 Conclusion

In this study, we analyzed GPCR assays from PUBCHEM with the aim to identify frequent hitters. We found that fluorescence-based assays are more susceptible to false positives than bioluminescence. Compounds that were frequent hitters at fluorescence-based assays did not appear as frequent hitters in bioluminescence assays. A predictive machine-learning model to identify such compounds for GPCR assays was developed. The provided analysis can help to interpret HTS screening using GPCR assays.

Acknowledgement. The project leading to this report has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676434, "Big Data in Chemistry". The article reflects only the author's view and neither the European Commission nor the Research Executive Agency (REA) are responsible for any use that may be made of the information it contains. We thank Michael Withnall for English correction. The authors thank ChemAxon (<http://www.chemaxon.com>) for Academic license of software tools (Standartizer, ChemAxon plugins) as well as AlvaScience (<http://alvascience.com>), Molecular Networks GmbH (<http://mn-am.com>) and Chemosophia (<http://chemosophia.com>) for providing descriptors and Corina 2D to 3D conversion program used in this study.

References

1. Hauser, A.S., Attwood, M.M., Rask-Andersen, M., Schioth, H.B., Gloriam, D.E.: Trends in GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discov.* **16**, 829–842 (2017). <https://doi.org/10.1038/nrd.2017.178>
2. Fredriksson, R., Lagerström, M.C., Lundin, L.-G., Schiöth, H.B.: The G-protein-coupled receptors in the human genome form five main families. phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **63**, 1256–1272 (2003). <https://doi.org/10.1124/mol.63.6.1256>
3. Zhang, R., Xie, X.: Tools for GPCR drug discovery. *Acta Pharmacol. Sin.* **33**, 372–384 (2012). <https://doi.org/10.1038/aps.2011.173>
4. Ghosh, D., Koch, U., Hadian, K., Sattler, M., Tetko, I.V.: Luciferase advisor: high-accuracy model to flag false positive hits in luciferase HTS assays. *J. Chem. Inf. Model.* **58**, 933–942 (2018). <https://doi.org/10.1021/acs.jcim.7b00574>
5. Schorpp, K., Rothenaigner, I., Salmina, E., Reinshagen, J., Low, T., et al.: Identification of small-molecule frequent hitters from AlphaScreen high-throughput screens. *J. Biomol. Screen.* **19**, 715–726 (2014). <https://doi.org/10.1177/1087057113516861>
6. Roche, O., Schneider, P., Zuegge, J., Guba, W., Kansy, M., et al.: Development of a virtual screening method for identification of “frequent hitters” in compound libraries. *J. Med. Chem.* **45**, 137–142 (2002). <https://doi.org/10.1021/jm010934d>
7. Hu, G.M., Mai, T.L., Chen, C.M.: Visualizing the GPCR network: classification and evolution. *Sci. Rep.* **7**, 15495 (2017). <https://doi.org/10.1038/s41598-017-15707-9>
8. Stevens, R.C., Cherezov, V., Katritch, V., Abagyan, R., Kuhn, P., et al.: The GPCR network: a large-scale collaboration to determine human GPCR structure and function. *Nat. Rev. Drug Discov.* **12**, 25 (2012). <https://doi.org/10.1038/nrd3859>
9. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., et al.: PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2018). <https://doi.org/10.1093/nar/gky1033>
10. Sushko, I., Novotarskyi, S., Körner, R., Pandey, A.K., Rupp, M., et al.: Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* **25**, 533–554 (2011). <https://doi.org/10.1007/s10822-011-9440-2>
11. Tetko, I.V.: Associative neural network. *Neural Process. Lett.* **16**, 187–199 (2002). <https://doi.org/10.1023/a:1019903710291>
12. Sosnin, S., Karlov, D., Tetko, I.V., Fedorov, M.V.: Comparative study of multitask toxicity modeling on a broad chemical space. *J. Chem. Inf. Model.* **59**, 1062–1072 (2019). <https://doi.org/10.1021/acs.jcim.8b00685>
13. Chen, T., Guestrin, C.: XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2016, San Francisco, California, USA, pp. 785–94. ACM (2016). <https://doi.org/10.1145/2939672.2939785>
14. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. *Neural Process. Lett.* **9**, 293–300 (1999). <https://doi.org/10.1023/a:1018628609742>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

